ASSUMPTIONS BEHIND LINEAR REGRESSION

SimpleAI.ch

DESCRIPTION

TESTING

HANDLING PROBLEMS

VALIDITY OF LR MODELS & LINEARITY

X ~Y should be a linear relationship

- *(good example: Anscombe Quartet*

and

y is a linear combination of the parm's (coefficients) and predictor var's (X).

Consequences: Xi are treated as fixed points, thus, only coeff's must be found + you can create many copies/transforms of X, as long as they are not exactly corr. with each other

If Problematic:

Linear model can not be used or it will make systematic, and biased results

Examine X~Y relationship

X~Y - IS IT LINEAR RELATIONSHIP?

- **Target variable vs predictor variable** ideally these should be linear

- Methods -

- *Scatterplots with smoothed line,*
- *Boxplots, violin plots, with trendline*

X~Y - ARE THEY CORRELATED?

- Calculate Correlation coeff for each predictor vs target var.

- Methods -

- *Pearson (linear), Sperman (rank), or kendal corr. Coef (rank pairs)*
- *Barplot with sorted corr. results,*
- *Table,*

Check Input data

ARE ALL FEATURES NORMALLY DISTR.?

- **Plots**; *histogram, boxplot, and QQ-plot*
- ***Descriptive;*** *kurtosis, skewness, mean*
- **Stats;** *most often: Shapiro–Wilk , Kolmogorov-Smirnov (>300 samples)*

Turn curved line of x~y into straight one &/or ensure that X values are norm. distr.

FEATURE TRANSFORMATIONS

Log, sqrt, power etc…

Not necessarily, to get norm distrib. values

ENSURE NORMALITY OF INPUT DATA

Especially important with >2 predictors

Fit model to a curved line of x~y

POLYNOMIAL REGRESSION

NON-LINEAR REGRESSION

NO MULTICOLINEARITY

Muticolinearity: two or more independent variables have a high correlation among themselves.

If Problematic:

- Computational problems: have difficulty in distinguishing between effects of correlated predictors on the dependent variable.
- ADD MORE DATA – it often helps

Examine Feature Values

FIND WHICH PAIRS OF PREDICTOR VARIABLES ARE CORRELATED WITH EACH OTHER

- ***Pair-wise comparisons (X)***

- Methods -

- ***Pearson*** *(linear),* ***Sp ….***
- ***Heatmap****, or table with top results (use absolute values, or head/tail)*

FIND WHICH PREDICTOR VARIABLE IS MOST LIKERY CORRELATED WITH THE REST OF THE DATA

- ***One-vs-rest comparisons (X)***

- Methods -

- **Variance Inflation Factor (VIF)** *one feature is used as dependent var. and all other features as its predictors*

Remove one column in each one-hot-encoded feature

Regularization

RIGDE /ELASTICNET or LARS REGR

Remove corr. features

REMOVE HIGHLY CORR. FEATURES *(LASSO, VIF>5, corr~-1/1)*

HIERARCHICAL CLUSTERING

Compute spearman rank order coeff. and pick a single feature from each cluster of features correlated with each other

Dimensionality Reduction

PCA PREPROCESSING

take the top eigenvectors that preserve the max. variance

NORMALLY DISTRIBUTED RESIDUALS

Required for validity of confidence intervals & model predictions

- In ideal situation, you would collect many Xi for each Y, and ideally, the measuring error that causes differences between Xi, would have normal distribution with mean=0.
- In most datasets, you don't have many Xi points for each Yi, but you can generalize, this assumption for all residuals calculated from entire Xi/Yi population.

Are residuals v. normally distributed?

VISUAL INSPECTION OF RESIDUALS

Part 1

- ***Histogram***
- ***QQ-Plots*** *(Quantilie-Quantile Plots), or* ***PP-Plots*** *(Probability Plots)*
- ***Plot residuals vs target variable*** *(look for bias above/below 0)*

DESCRITIVE STATISTICS

- ***Mean;*** *should be 0*
- ***Skewness;*** *should be 0*
- ***Kurtosis;*** *ideally ~2.2*

STATISTICAL TEST FOR NORMALITY

- ***Shapiro–Wilk test*** *(small and medium size datasets)*
- **Scipy normaltest** (up to~50 samples);
- **Lilliefors-test** (50-300 samples),
- ***Kolmogorov-Smirnov*** *(>300 samples)*

Try it in this order:

CHECK INPUT DATA NORMALITY

Use the same methods as in the above

IF NOT TRANSFORM THE DATA

Log, sqrt, repricotal, power etc…

TEST LINEARITY ASSUMPT. & ACT IF IT IS NOT HOLDING

See actions for the first assumption

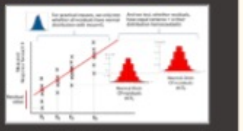
APPLY NON-LINEAR TRANFORMERS TO IMPUT DATA

- Quantile transformer
- Power transformer
  - 'box-cox' – positive data only
  - 'Yeo-Johnson' - ± data

HOMOSCEDASITY OF RESIDUALS

Homogeneity in variance of residuals = Equal Variance of Errors

often in time series data



If Problematic:

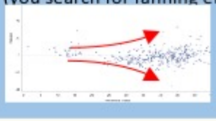
- regression prediction remains unbiased and consistent but inefficient - distinguishable variances are averaged to get a single variance that is inaccurately representing all the variances of the line (eg. small values have the same variance as large ones)

Are residuals values homoscedastic?

VISUAL INSPECTION OF RESIDUALS

Part 2

- ***Plot residuals vs target variable*** (vou search for fanning effect)



STATISTICAL TEST FOR HOMOSCEDACITY

- **White test**; detects, linear hetero-scedastity *Looses sensitivity, with large number of predictor variables*
- **Breusch-Pagan test** (more general, less sensitive)

Both tests works by creating an auxiliary regression, on the residuals, vs independent and dependent variables

Y transformation

TRANSFORM DEPENEDENT VARIABLE (Y)

Log, sqrt, box-cox transformed, or power transformer

Modify or use different cost function

ADD WEIGHTS TO ERROR TERM

Weighted OLS LR

MODIFY ERROR FUCNTION

Eg. Huber loss, for dealing with outliers

NO AUTO-CORRELATION

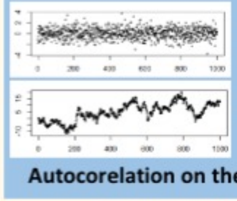
The correlation of a time series with its own past or the future values causes autocorrelation.

autocorrelation does not bias the OLS coefficient estimates. However, the standard errors tend to be underestimated

Are residuals randomly distributed?

Plot Residuals vs

Sample Order, Time variable, any spatial variable, or any other variable, used in the model



No autocorelation *random assortment of residuals*

Autocorelation on the plot

DURBIN -WHATSON TEST

- measures the amount of autocorrelation in residuals from the regression analysis.
- check sONLY for the first-order autocorrelation, ie. lag =1

IMPROVE THE MODEL

- Add new spatial/timedependent variables
- Tune hyperparameters
- Use Autoregressive models (AR1 model)

FEATURE TRANNSFORMATION

Transforming var's & test if the autocorrelations were reduced.

- deviation from the average values; eg: weather data.
- Log or exponential - may or may not make improvements.
- Annualising - to remove seasonal eff

CLUESTERING

Clustering on t time invariant factors.

ASSUMPTIONS BEHIND LINEAR REGRESSION – PART 1

SimpleAI.ch

LINEAR REGRESSION ASSUMPTIONS

1. *Validity of Linear Models*
2. *No multicolinearity in the data*
3. *Normally distributed residuals*
4. *Homoscedasticy of residuals / constant variance*
5. *No auto-correlation*

WHICH ASSUMPTION IS THE MOST IMPORTANT

First, check for validity of linear models, next independence assumptions, such as multicollinearity, and autocorrelation. Subsequently, test any equal variance assumption, and finally the assumptions on distribution (e.g., normal) - Techniques are usually least robust to departures from independence and most robust to departures from normality

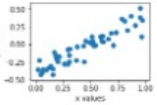
Robustness to departures from normality is related to the Central Limit Theorem, since most estimators are linear combinations of the observations, and hence approximately normal if the number of observations is large

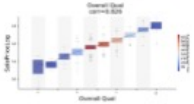
If a linear model fits with all predictors included, it is *not* true that a linear model will still fit when some predictors are dropped. For example, if $E(Y|X_1, X_2) = 1 + 2X_1 + 3X_2$ (so that a linear model fits when Y is regressed on both X_1 and X_2), but $E(X_1|X_2) = \log(X_1)$, then it can be calculated that $E(Y|X_1) = 1 + 2X_1 + 3\log(X_1)$, which says that a linear model does not fit when y is regressed on X_1 alone.

VALIDITY OF LR MODELS & LINEARITY ASSUMPTION

HOW TO TEST IT?

Visual examination of x~y

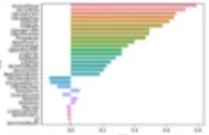




***linearity assumption** - the conditional means of the response variable are a linear function of the predictor variable. Graphing the response variable vs the predictor can often give a good idea of whether or not this is true. However, one or both of the following refinements may be needed:*

- **Plot residuals (instead of response) vs. predictor.** A non-random pattern suggests that a simple linear model is not appropriate; you may need to transform the response or predictor, or add a quadratic or higher term to the mode.
- **Use a scatterplot smoother such as lowess** (also known as loess) to give a visual estimation of the conditional mean. Such smoothers are available in many regression software packages. *Caution*: You may need to choose a value of a smoothness parameter. Making it too large will oversmooth; making it too small will not smooth enough

Correlation analysis



Eg: Use Pearson's correlation, but, remember to Check coefficient value and the significance of the correlation (p-value) - p-value is indicating that there is an absence or presence of a Significant relationship between variables

SOURCES

Much of the text on this slide, and the figure taken from:
https://medium.com/geekculture/holy-grail-for-understanding-all-the-assumptions-of-linear-regression-[illegible]
https://www.geeksforgeeks.org/detecting-multicollinearity-with-vif-python/
https://stackoverflow.com/questions/[illegible]/variance-inflation-factor-in-python
https://towardsdatascience.com/7-techniques-to-handle-multicollinearity-that-every-data-scientist-should-know-[illegible]
https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html#handling-multicollinear-features
https://web.ma.utexas.edu/users/mks/statmistakes/modelcheckingplots.html

Dealing with muticolinearity - examples
https://scikit-learn.org/stable/auto_examples/inspection/plot_permutation_importance_multicollinear.html
https://github.com/bhattbhavesh91/pca-multicollinearity/blob/master/multi-collinearity-pca-notebook.ipynb

HANDLING

FIRST:
We try to turn curved line of x~y relationship into straight one, and to apply linear regression model s – these are easier

SOLUTION 1. TRY FEATURE TRANSFORMATIONS

- **Scaling & shift (f(x)±c) –not very helpful,**
- **Reflection** (-f(x), shifts values on the other site of the axis.
- **log transformation** (log(x), ln(x), used for right-skewed distributions),
  - Data skewed to the right (i.e. in the positive direction).
  - The residual's standard deviation is proportional to fitted values
  - The data's relationship is close to an exponential model
  - the residuals reflect multiplicative errors that have accumulated during each step of the computation
- **Sqrt transformations;** it basically makes a straight line from power fn line, and its used. Sqrt transf. compresses larger values, and makes differences between small values more apparent, but it is made less aggressively then log transf.
- **power transformation** (eg: x^2), used when The data's relationship is close to an exponential model
- **repricotal transformation** (1/x, - dramatic effect on the shape of the distribution, reversing the order of values with the same sign. The transformation can only be used for non-zero values.),
- **Share mapping** (all points along one line stay fixed, while other points are shifted parallel to the line by a distance proportional to their perpendicular distance from the line)
- More info fromhttps://www.calculushowto.com/transformations/

THEN:
If solution 1 fails, we must fit the model to curved x~y line

SOLUTION 2. USE POLYNOMIAL REGRESSION

Allow for the polynomial linear regression equation. While the independent variable is raised to power of 2 here, the model is still linear in terms of its parameters. Linear models can also contain log terms and inverse terms to follow different kinds of curves and yet continue to be linear in the parameters

SOLUTION 3. APPLY NON-LINEAR REGRESSION MODEL

CAUTION:
It is not possible to gauge from scatterplots whether a linear model in more than two predictors is suitable.

TRANSFORM PREDICTORS TO APPROX. MUTIVARIATE NORMALITY

Ii will ensure not only that a linear model is appropriate for all (transformed) predictors together, but that a linear model is appropriate even when some transformed predictors are dropped from the model.

multivariate normality

multivariate normal distribution, multivariate Gaussian distribution, or **joint normal distribution** is a generalization of the one-dimensional (univariate) normal distribution to higher dimensions. One definition is that a random vector is said to be *k*-variate normally distributed if every linear combination of its *k* components has a univariate normal distribution.

Bivariate normal joint density

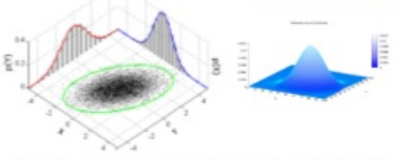


Illustration of a multivariate gaussian distribution and its marginals.

NO MULTICOLINEARITY IN INPUT DATA

COLINEAR FEATURES

exact linear relationship among two or more independent variables

- *Eg: weight in kg. and pounds*
- **Collinearity** causes problems in numerical computations. Eg: alg tries to decide which of the same variables are more important
- No solution for **OLS method**, == all features must be linearly independent.
- **Most alg.** can cope with that, but they will return warnings
- Collinearity is can be introduced by poor design & must be detected / handled before building a model

NEAR COLINEARITY & ILL-CONDITIONING

Nearly Col. F's, have "imperfect" or close to linear relationship

- *Eg: weight from 2 diff. balances*
- **OLS solution exist**, but the calculations are unstable
- **Ill-Conditioning:** matrix with nearly colinear features, has large condition nr that is used to find OLS solution, and it is susceptible to small changes in input data. Thus it causes large variation in model coeff. with small changes in the data
- In most cases, ill-conditioning doesn't affects model accuracy!

HOW TO DETECT MUTICOLINEARITY ?

Correlation analysis

- pair-wise comparisons only

Variance Inflation Factor (ViF)

- Allows estimating multicollinearity in entire dataset, vs each feature,
- The value of VIF is computed for each feature, where a regression model is trained keeping one feature as dependent var. and all other features as its predictors

$$VIF_i = \frac{1}{1 - R_i^2}$$

If R^2 ->1**, then the regr. Valuesin the i-th feature can be easily predicted with the rest of the data == potential mucticolinearity, **(VIF >>1)

If R^2 -> 0**, then LR model is "bad" and there is proof for collinearity between i-th feature and the rest of the dataset **(VIF ->1)

Values = [0, inf)
- VIF=1: No multicollinearity
- VIF 1 - 5: Moderate multicollinearity
- VIF > 5: Highly multicollinear

HANDLING

SOLUTION 1. REMOVE HIGHLY CORR. FEATURES

- LASSO REGRESSION
- REMOVE HIGHLY CORRELATED FEATURES
  foe example: With Vif >= 5 /or/. Corr ~-1 or 1

SOLUTION 2. USE RIGDE /ELASTICNET or LARS REGR

Ridge — used for data with large # of predictors of about the same value

ElasticNet — Lasso is likely to pick one of correlated features at random, while elastic-net is likely to pick both, but with smaller coefficients,

LARS — alternative to ridge & ElasticNet for multiple colinear features, iterative approach, **More effective for p>>n datasets, and alfa exploration**

SOLUTION 3. DIMENSIONALITY REDUCTION : PCA PREPROCESSING

taking the top eigenvectors that preserve the maximum variance. The number of dimensions can be decided by observing the variance preserved for each eigenvector.
https://github.com/bhattbhavesh91/pca-multicollinearity/blob/master/multi-collinearity-pca-notebook.ipynb

SOLUTION 4. HIERARCHICAL CLUSTERING

Perform hierarchical clustering on the spearman rank order coefficient and pick a single feature from each cluster based on a threshold. The threshold can be decided by observing the dendrogram plots.

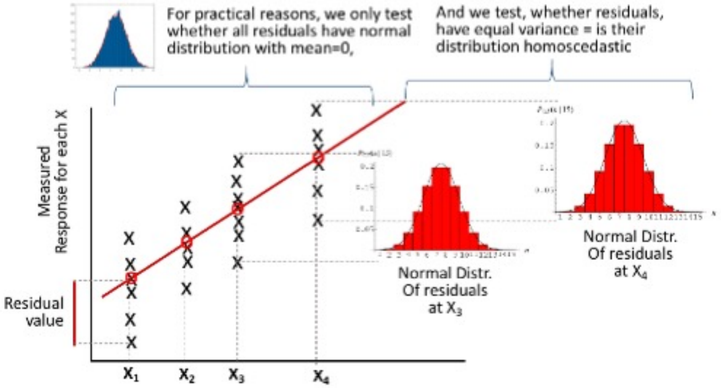
ASSUMPTIONS BEHIND LINEAR REGRESSION – PART 2

SimpleAI.ch

LINEAR REGRESSION ASSUMPTIONS

1. *Validity of Linear Models*
2. *No multicolinearity in the data*
3. *Normally distributed residuals*
4. *Homoscedasticy of residuals / constant variance*
5. *No auto-correlation*

NORMAL DISRT. OF RESIDUALS/ERRORS


For practical reasons, we only test whether all residuals have normal distribution with mean=0,

And we test, whether residuals, have equal variance = is their distribution homoscedastic

Measured Response for each X
Residual value
X_1
X_2
X_3
X_4
Normal Distr. Of residuals at X_3
Normal Distr. Of residuals at X_4

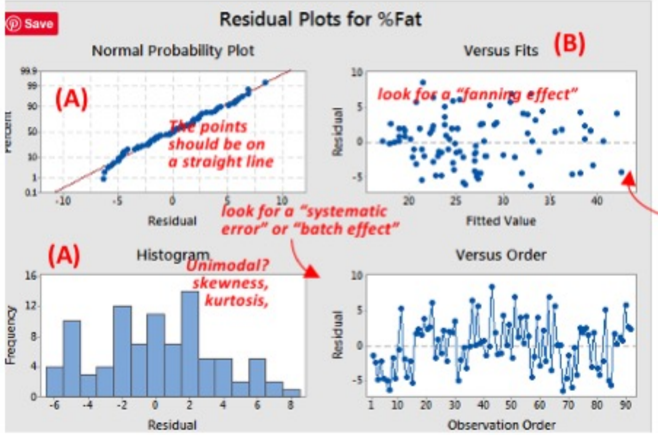
Note on Heteroscedasticity

It will result in the averaging over of distinguishable variances around the points to get a single variance that is inaccurately representing all the variances of the line. In effect, residuals appear clustered and spread apart on their predicted plots for larger and smaller values for points along the linear regression line, and the mean squared error for the model will be wrong.

Caution: Hypothesis tests for equality of variance are often not reliable, since they also have model assumptions and are typically not robust to departures from these assumptions.

CHECK DISTRIBUTION OF RESIDUALS ON PLOTS

(A) Use: Quantilie-Quantile Plots (QQ-Plots), PP-Plots (Probability Plots) the goal is to check if the points from two distributions lie on the y~x line.


Residual Plots for %Fat

(A) The points should be on a straight line

(B) look for a "fanning effect"

(A) Unimodal? skewness, kurtosis,

look for a "systematic error" or "batch effect"

Use Statistical tests for normality

Small sample-numbers (<50)

- **Shapiro–Wilk W test;** depends on the cov. matrix between the order statistics of the observations – less sensitive to outliers then normal test
- **scipy normaltest;** combines skew and kurtosis

Intermediate sample numbers (50-300)

- **Shapiro–Wilk W or Lilliefors-test**
- **Lilliefors-test;** based on Kolmogorov–Smirnov test, - quantifies distance between empirical distr. fn. of the sample and the cdf of the reference distrib (eg normal distrib), or between dictrib. Fn's of two samples. It is good because original K-S-test is unreliable when mean and std are unknown.

large sample numbers (>300)

- While having sufficient sample size, **Kolmogorov-Smirnov** and **Lilliefors-test**, are the least affected with extreme values (outliers), followed with **Shapiro–Wilk test**, that is more affected, but less then the **normaltest**

HANDLING

SOLUTION 1.

TEST LINEARITY ASSUMPTION & ACT IF IT IS NOT ….

First check if the linearity assumption is being violated. That can cause a failure with the normality assumption as well.

SOLUTION 2.

CHECK INPUT DATA NORMALITY & TRANSFORM THEM TO HAVE IT

Perform univariate analysis of dependent & independent variables and see **if the are significantly deviating from the normal distribution (see statistical tests for normality in the above)**. In this you can transform features (log, sqrt, power, or repricotal etc.)

SOLUTION 3.

APPLY NON-LINEAR TRANFORMERS TO IMPUT DATA

- **Quantile transformer**
- **Power transformer**
  - **'box-cox'** - needs the data to be positive
  - **'Yeo-Johnson'** - data to be both negative and positive.

SOURCES

https://realpython.com/numpy-scipy-pandas-correlation-python/#rank-correlation
https://docs.tibco.com/data-science/GUID-75AB887A-9927-4772-BC83-DA09E65B3387.html
https://statisticsbyjim.com/regression/variance-inflation-factors/
https://medium.com/geekculture/holy-grail-for-understanding-all-the-assumptions-of-linear-regression-210f224192b5
https://stackoverflow.com/questions/42658379/variance-inflation-factor-in-python

RESIDUALS HOMOSCEDATICITY

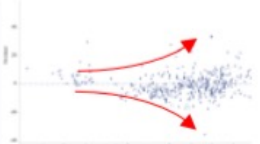
Homogeneity in variance of the residuals
= Equal Variance of Errors

CONSEQUENCES*: The regression prediction remains unbiased and consistent but inefficient. It is inefficient because the estimators are no longer the Best Linear Unbiased Estimators (BLUE). The hypothesis tests (t-test and F-test) are no longer valid.*

HOW TO TEST IT?

Plot Residuals vs Fitted Values (Plot B)

Look for **fanning effect** on the scatterplot with residuals vs the dependent variable



Statistical tests Homoscedastic

- **White test:** *>>> statsmodels.stats.diagnostic.het_white*
- **Breusch-Pagan test** *>>> het_breuschpagan in statsmodels.stats.diagnostic*

SIMILARITY BETWEEN THESE TWO TESTS

- Both tests works, similar by creating an auxiliary regression, on the residuals, vs independent and dependent variables
- The errors are heteroscedastic when p>0.05

& DIFFERENCES

- The **Breusch-Pagan** test only checks for the linear form of heteroskedasticity
- The **White test** is more generic. It relies on the intuition that if there is no heteroskedasticity the classical error variance estimator should gives you standard error estimates close enough to those estimated by the robust estimator (based on median). A shortcoming of the White test is that it can lose its power very quickly particularly if the model has many regressors.

HANDLING

SOLUTION 1. TRANSFORM DEPENEDENT VARIABLE

Typically we use log, transformation, or box-cox tranformations (power transformer for positive only, values) Comment: it often happens in time series data, caused by season, monthly, and other patterns

SOLUTION 2. ADD WEIGHTS TO ERROR TERM -

eg use Weighted OLS LR

For example use weighted least squares in case all the transformations fail to solve the problem at hand. LinearRegressor, in Sklear, takes weights as the third parameter after X, y, and the weights are multiplied by errors, calculated from each data point – hence scaling of weights doesn't change anything.

How to construct weights

1. Compute the absolute and squared residuals
2. Find the absolute and squared residuals vs. independent variables to get the estimated standard deviation and variance
3. Compute the weights using the estimated standard deviations & variance.

SOLUTION 3. MODIFY ERROR FUCNTION

eg use Huber Regression

ASSUMPTIONS BEHIND LINEAR REGRESSION – PART 3

SimpleAI.ch

NO AUTO-CORRELATION
- Independence of errors -

AUTOCORRELATION

A measure of similarity between a given time series and the lagged version of the same time series over successive time periods. In other words: It is a correlation between two different versions X_t and X_{t-k} of the same time series.

Partial autocorrelation function PACF

- PACF gives the partial correlation of a stationary time series with its own lagged values, regressed the values of the time series at all shorter lags. It is different from the autocorrelation function, which does not control other lags then 1.

Causes

- The correlation of a time series with its own past or the future values causes autocorrelation. Generally, any usage has a tendency to remain in the same state from one observation to the next. This specific form of 'persistence' causes the positive autocorrelation.

Effects

- autocorrelation it does not bias the OLS coefficient estimates. However, the standard errors tend to be underestimated

Usage

- **For checking randomness in the time-series**; In many statistical processes, our assumption is that the data generated is random (autocorrelation of lag 1)
- **To determine whether there is a relation between past and future values** of time series, we try to lag between different values.

Challenge with testing independence assumptions

Ind. Ass. are usually formulated in terms of error terms rather than in terms of the outcome variables. For example, in simple linear regression, the model equation is Y = α + βx + ε, where Y is the outcome (response) variable and ε denotes the error term (also a random variable).

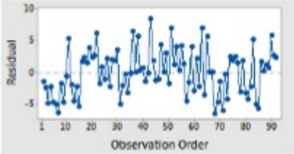
It is the error terms that are assumed to be independent, not the values of the response variable. We do not know the values of the error terms ε, so we can only plot the residuals e_i (defined as the observed value y_i minus the fitted value, according to the model), which approximate the error terms.

HOW TO TEST IT?

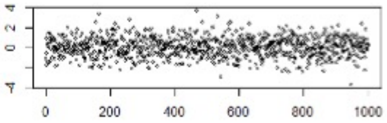
CHECK RESIDUALS

Plot Residuals vs

- Sample Order
- any Time variable
- any spatial variable
- any variable, used in the model

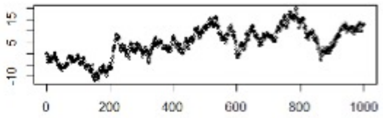


A pattern that is not random suggests lack of independence.



No autocorelation

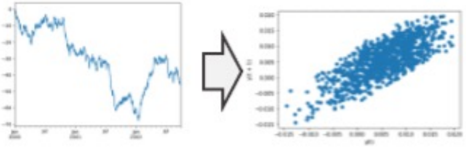
random assortment of residuals



Autocorelation on the plot

If we compare each value to their preciding/subsequent values in order on X axis, high values will be often correlated with similarly high values, and low values with similarly low values, irrespectively on the pattern

```
>>> pd.plotting.lag_plot(df, lag = 1)
```



Autocorelation plot can be done with pandas function,

other plots are also available for deeper anaylsis

Statistical Test For Autocorrelation

Durbin-Watson Test:

Used to measure the amount of autocorrelation in residuals from the regression analysis.
It is used to check ONLY for the first-order autocorrelation, ie. lag =1

Assumptions

- The errors are normally distributed and the mean is 0.
- The errors are stationary.

Formula

$$DW = \frac{\sum_{t=2}^{T}(e_t - e_{t-1})^2}{\sum_{t=1}^{T} e_t^2}$$

e_t - residual of error from the Ordinary Least Squares (OLS) method

Values

The Durbin Watson test has values between 0 and 4.

- 2: **No autocorrelation. Generally, we assume 1.5 to 2.5 as no correlation.**
- **0- <2: positive autocorrelation.** The more close it to 0, the more signs of positive autocorrelation.
- **>2 -4: negative autocorrelation.** The more close it to 4, the more signs of negative autocorrelation.

CAUTION

These test, have problems with detecting many types of autocorrelation, thus, visual inspection is always recommended + you can perform more plots to support your claims for autocorrelation at different lag values

HANDLING autocorrelation

IMPROVE THE MODEL

Try to capture structure in the data in the model. ...

- Adding other variables as independent variables
- Experimenting with model specification
- Use Autoregressive model (AR1 model)

FEATURE TRANNSFORMATION

- Transforming variables into different functional forms;
- Many variables can be transformed into different forms and tested to see if the autocorrelations were reduced. (log, exponential, annulized etc...)
- Examples:
  - deviation from the average values; eg: weather data.
  - Log form or exponential form may or may not make improvements.
  - Annualising the data - it is supposed to remove any seasonal effects.

CLUESTERING

- Clustering on different time invariant factors
- Clustering based on variables, such as income and lot size, can improve the autocorrelations problems. This may be due to the phenomenon of seasonality, with houses reacting differently to the same weather conditions.
- Caution. p values of estimates can get worse as the number of households within each cluster is smaller than the whole segment. Hence, the clustering analysis can be used to identify outliers and appropriate actions can be taken.

SOURCES
https://kumaradevan.com/how-to-handle-autocorrelation/
https://stats.stackexchange.com/questions/14914/how-to-test-the-autocorrelation-of-the-residuals
https://stats.stackexchange.com/questions/50151/how-to-tell-if-residuals-are-autocorrelated-from-a-graphic?noredirect=1&lq=1
https://www.geeksforgeeks.org/autocorrelation/

REGRESSION ANALYSIS

COREELARTION COEFFICIENTS

SimpleAI.ch

CORRELATION TYPES

There are several types of correlation metrics that can be used for different types of data: The most popular are

- **Pearson's coefficient** that measures linear correlation in numerical data
- **and Spearman or Kendall coefficients used to** compare the ranks of data

Correlation Technique	Relationship	Dataype of features
Pearson	Linear	Quantitative and Quantitative
Spearman	Non-linear	Ordinal and Ordinal
Point-biserial	Linear	Binary and Quantitative
Cramer's V	Non-linear	Categorical and Categorical
Kendall's tau	Non-linear	Two Categorical or Two Quantitative

SciPy CODE EXAMPLES

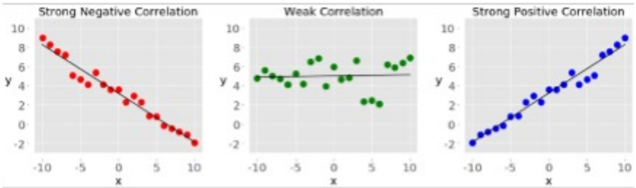
All Scipy Functions Return
-> corr. coefficient
-> *p-value*

```
# create example data
>>> x = np.arange(10, 20)
>>> y = np.array([2, 1, 4, 5, 8, 12, 18, 25, 96, 48])

# Pearson's r
>>> scipy.stats.pearsonr(x, y)
(0.7586402890911869, 0.010964341301680832)

# Spearman's rho
>>> scipy.stats.spearmanr(x, y)
SpearmanrResult( correlation=0.975757, pvalue=1.4675461877e-06)

# Kendall's tau
>>> scipy.stats.kendalltau(x, y)
KendalltauResult(correlation=0.911111, pvalue=2.976190462e-05)
```

CAUTION

Check NA policy in each function

INTERPRETATION



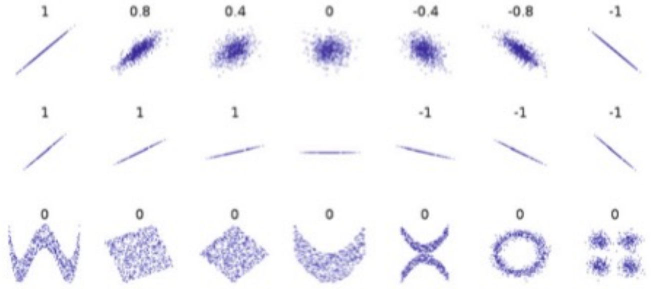


Fig. 11.1 Several sets of (x, y) points, with the correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (*top row*), but not the slope of that relationship (*middle*), nor many aspects of nonlinear relationships (*bottom*). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero. (In Wikipedia. Retrieved May 27, 2015, from http://en.wikipedia.org/wiki/Correlation_and_dependence.)

PEARSON R *CORRELATION COEEFICIENT*

The *correlation coefficient* between two variables answers the question: "Are the two variables related? That is, if one variable changes, does the other also change?" If the two variables are normally distributed, the standard measure of determining the correlation coefficient, often ascribed to Pearson, is

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad (11.1)$$

With the sample covariance s_{xy} defined as

$$s_{xy} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \qquad (11.2)$$

and s_x, s_y the sample standard deviations of the x and y values, respectively, Eq. 11.1 can also be written as

$$r = \frac{s_{xy}}{s_x \cdot s_y}. \qquad (11.3)$$

Pearson's correlation coefficient, sometimes also referred to as *population correlation coefficient* or *sample correlation*, can take any value from −1 to +1. Examples are given in Fig. 11.1. Note that the formula for the correlation coefficient is symmetrical between x and y—which is not the case for linear regression!

REPORTING RESULTS

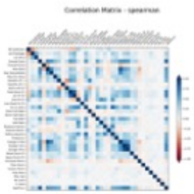
Reporting individual Results:

- ... and ... were found to be moderately positively corr., r(38) = .34, p = .032.
- .... were found to be strongly correlated, r(128) = .89, p < .01.
- .... were negatively correlated, r(78) = -.45, p < .001
- no linear correlation was found, between ..&..., r(38) = .02, p = .005

Comments

- Degrees of freedom for r is N - 2 – denoted in brackets -> r(120)
- Report the exact *p*value, + state your alpha, eg 0.05 level early in your results
- r statistic should be stated at 2, or 3 decimal places

CORRELATION MATRIX & TRENDLINE



Pandas

```
>>> df_corr = df.corr(method='pearson')
>>> fig, ax = plt.subplots()
>>> ax.matshow(
            df_corr, cmap=cmap, vmin= -1, vmax=1)
    # min,max values will scale the heatmap
```

Scipy

```
>>> rho, pvalues = scipy.stats.spearmanr(np.array)
```

Obtained with linear regr. In scipy

```
>>> result = scipy.stats.linregress(x, y)
    result.slope          # 7.4363636363636365
    result.intercept      # -85.92727272727274
    result.rvalue         # 0.7586402890911869
    result.pvalue         # 0.010964341301680825
    result.stderr         # 2.257878767543913
```

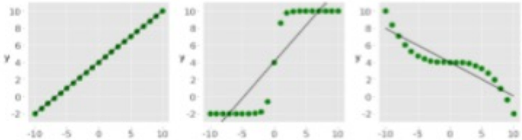
RANK CORRELATION

- ***For Not normally distributed data***
- ***For Non-linear relationships between compared var's***

SPEARMAN RHO

The **Spearman correlation coefficient** Is calculated the same way as the Pearson correlation coefficient but takes into account their ranks instead of their values.

- Denoted with the Greek letter rho (ρ) and called **Spearman's rho**.
- −1 ≤ ρ ≤ 1.
    - If ρ=1 => the function between x and y increases monotonically
    - If ρ=1 => the function between x and y decreases monotonically



Rank correlation, allows comparing variables with non-linear relationships

Sopearman Rho uses the ranks or the orderings of data points in compared variables/features, If the orderings are similar, then the correlation is strong, positive, and high. However, if the orderings are close to reversed, then the correlation is strong, negative, and low. In other words, rank correlation is concerned only with the order of values, not with the particular values from the dataset.

KENDAL TAU

Kendal Tau CC is harder to calculate than Spearman's but its confidence intervals are more reliable

- **Kendall correlation coefficient** is calculated as the difference in the counts of concordant and discordant ranked data pairs, relative to the total number of x-y pairs in compared rankings.
- VALUES:
    - −1 ≤ τ ≤ 1.
    - τ = 1 - the ranks of the corresponding values in **x** and **y** are the same. In other words, all pairs are concordant.
    - τ = −1 - the rankings in **x** are the reverse of the rankings in **y**. In other words, all pairs are discordant.

HOW IT IS DONE

- Lets, have two random variables x, & y, that we wish to compare
- All values in x, and y were ranked independently.
- Now, if we compare all pairs of values in x, and y, using the same pairs in xi:xj, and yi:xj in both variables at each comparison, we will label the results in one the 3 following classes:
    - **Pair of points has concordant ranks in x,y;** *($x_i > x_j$ and $y_i > y_j$) or ($x_i < x_j$ and $y_i < y_j$)*
    - **... has discordant** ranks in x,& y ; *($x_i < x_j$ and $y_i > y_j$) or ($x_i > x_j$ and $y_i < y_j$)*
    - **... the same ranks in x&y (tie);** *a tie in x ($x_i = x_j$) or a tie in y ($y_i = y_j$), or in both,*

According to the `scipy.stats` official docs, the Kendall correlation coefficient is calculated as τ = (n⁺ − n⁻) / √((n⁺ + n⁻ + nˣ)(n⁺ + n⁻ + nʸ)), where:

- n^+ is the number of concordant pairs
- n^- is the number of discordant pairs
- n^x is the number of ties only in **x**
- n^y is the number of ties only in **y**

If a tie occurs in both **x** and **y**, then it's not included in either n^x or n^y.

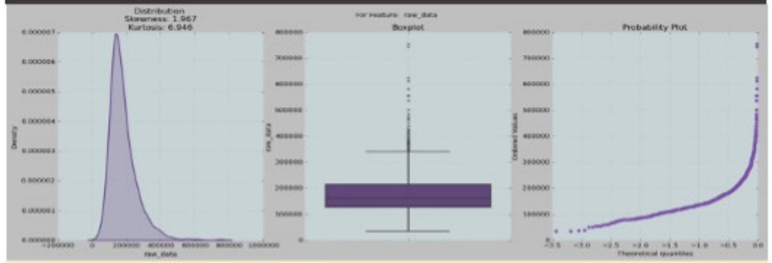
Greek letter tau (τ)

KENDAL TAU – DIFFERENT TYPES

- **scipy.stats.kendalltau has a.&b, varinats, that treat ties differently, b is default**
- **additionally, scipy offers weighted kendal tau,** in which exchanges of high weight are more influential than exchanges of low weight.

TESTING NORMALITY

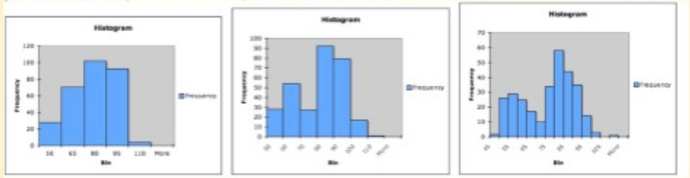
VISUAL EXAMINATION



Histogram vs Probability Plots

A histogram (whether of outcome values or of residuals) is *not* a good way to check for normality, since histograms of the same data but using different bin sizes (class-widths) and/or different cut-points between the bins may look quite different (see example below.

>> Instead, use a *probability plot* (also know as a *quantile plot* or *Q-Q plot*). *Caution*: Probability plots for small data sets are often misleading; it is very hard to tell whether or not a small data set comes from a particular distribution.



INTEPRETATION

(a) Two Identical distributions:

- the Q–Q plot follows the 45° line y = x.

(b) Two distributions agree after linearly transforming the values in one of the distributions,

- Q–Q plot follows some line, but not necessarily the line y = x.

(c) the distribution plotted on the x-axis is more dispersed than the distribution plotted on the vertical axis.

- the general trend of the Q–Q plot is flatter than the line y = x,

(d) the distribution plotted on the y-axis is more dispersed than the distribution plotted on the horizontal axis.

- Q–Q plot is steeper than the line y = x

(e) one of the distributions is more skewed than the other, or that one of the distributions has heavier tails than the other.

- Q–Q plots are often arced, or "S" shaped

Although a Q–Q plot is based on quantiles, in a standard Q–Q plot it is not possible to determine which point in the Q–Q plot determines a given quantile. For example, it is not possible to determine the median of either of the two distributions being compared by inspecting the Q–Q plot. Some Q–Q plots indicate the deciles to make determinations such as this possible.

PROBABILITY PLOTS

DEF: graphical technique for assessing whether or not a data set follows a given distribution such as the normal or Weibull. The data are plotted against a theoretical distribution in such a way that the points should form approximately a straight line.

1. P-P PLOT

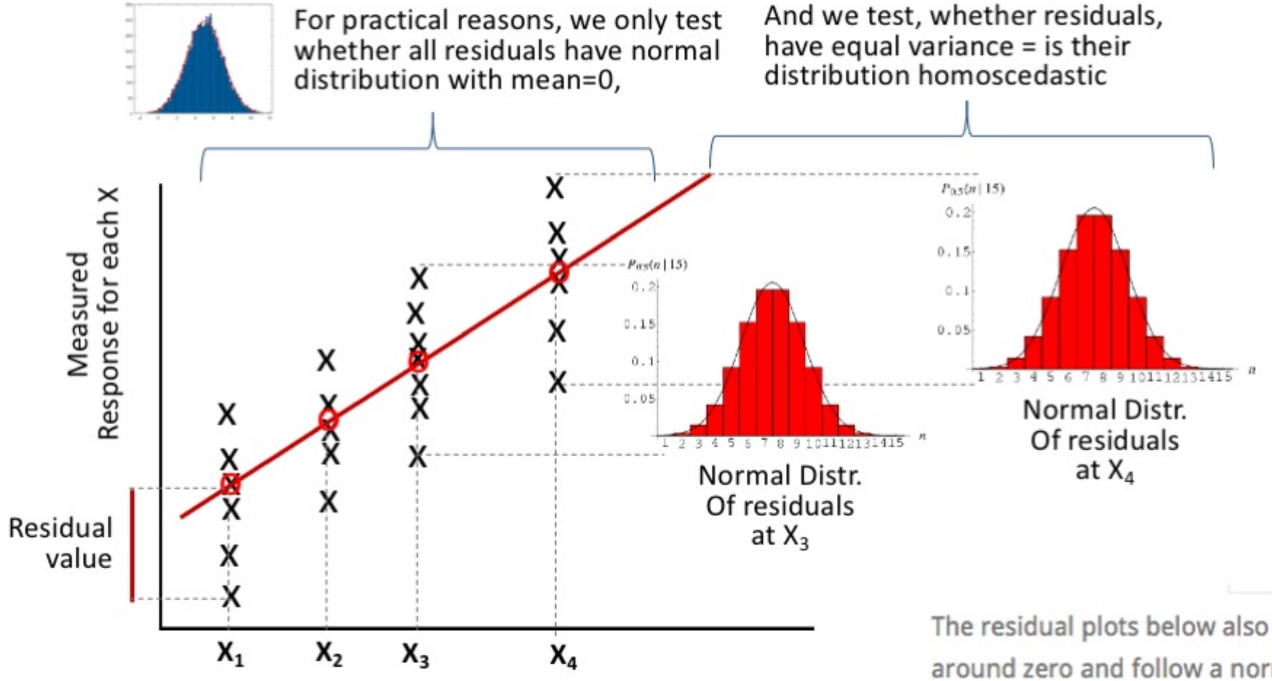
- "Probability-Probability" or "Percent-Percent" plot
- USED
  - to asses skenwness of the distribution
- LIMITATION: it is only useful for comparing probability distributions that have nearby or equal location. if two distributions are separated in space, the P–P plot will give very little data

2. Q-Q PLOT

- "Quantile-Quantile" plot
- **NORMAL PROBABILITY PLOT -** a Q–Q plot against the standard normal distribution
- DEF:
  - A Q–Q plot is a plot of the quantiles of two distributions against each other, or a plot based on estimates of the quantiles. The pattern of points in the plot is used to compare the two distributions.
  - Non-parametric approach to compare two datasets distribusiosn
- USED
  - plots the two cumulative distribution functions against each other.
  - compare two theoretical distributions to each other.
  - Or to compare collections of data, or theoretical distributions, eg. normal distribution
- MAIN ADVANTAGE
  - more widely used then PP-plots
  - Since Q–Q plots compare distributions, there is no need for the values to be observed as pairs, as in a scatter plot, or even for the numbers of values in the two groups being compared to be equal.
- RESULTS PROVIDED
  - Q-Q plot allows comparing the shapes of distributions, providing a graphical view of location, scale, skewness are similar or different in the two distributions.

https://realpython.com/numpy-scipy-pandas-correlation-python/#rank-correlation
https://docs.tibco.com/data-science/GUID-75AB887A-9927-4772-BC83-DA09E6583387.html
https://statisticsbyjim.com/regression/variance-inflation-factors/
https://medium.com/geekculture/holy-grail-for-understanding-all-the-assumptions-of-linear-regression-210f224192b5
https://stackoverflow.com/questions/42658379/variance-inflation-factor-in-python

Analysis of residuals



For practical reasons, we only test whether all residuals have normal distribution with mean=0,

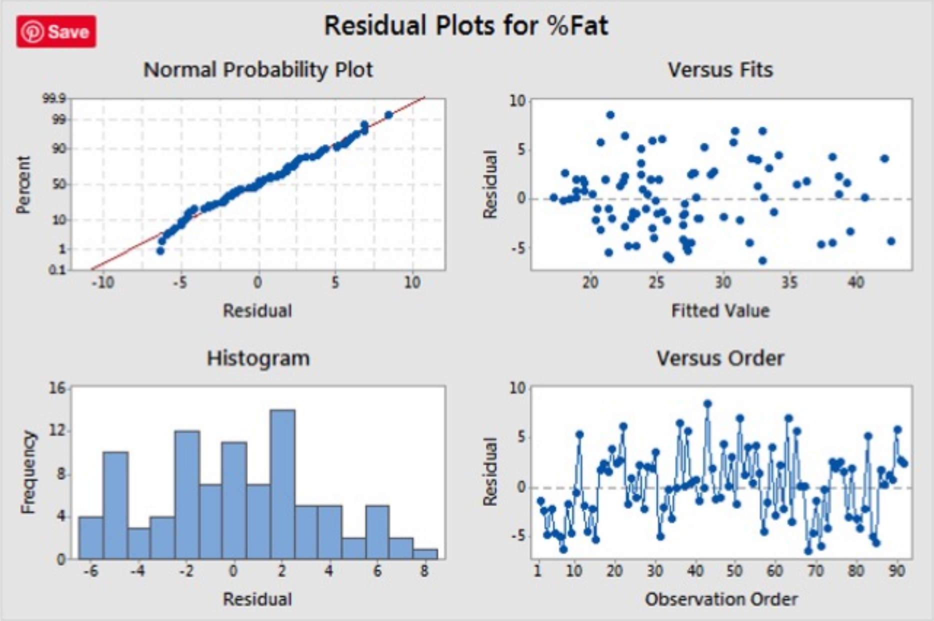
And we test, whether residuals, have equal variance = is their distribution homoscedastic

Measured Response for each X
Residual value
X_1
X_2
X_3
X_4
$P_{xs}(n|15)$
0.2
0.15
0.1
0.05
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
n
Normal Distr. Of residuals at X_3
Normal Distr. Of residuals at X_4

1. Independence of errors from the values of the independent variables.
2. Independence of the independent variables.

EFFECT: Heteroscedasticity will result in the averaging over of distinguishable variances around the points to get a single variance that is inaccurately representing all the variances of the line. In effect, residuals appear clustered and spread apart on their predicted plots for larger and smaller values for points along the linear regression line, and the mean squared error for the model will be wrong.

The residual plots below also confirm the unbiased fit because the data points fall randomly around zero and follow a normal distribution.



Residual Plots for %Fat